UNIVERSITI PUTRA MALAYSIA

ATTRIBUTE SET WEIGHTING AND DECOMPOSITION APPROACHES FOR REDUCT COMPUTATION

QASEM AHMAD AL-RADAIDEH.

FSKTM 2005 7

ATTRIBUTE SET WEIGHTING AND DECOMPOSITION APPROACHES FOR REDUCT COMPUTATION

By

QASEM AHMAD AL-RADAIDEH

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

July 2005

DEDICATION

To the memory of my Father,

To my great Mother,

To my Wife and Daughters,

To my Brother and Sisters.

Qasem

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirements for the degree of Doctor of Philosophy

ATTRIBUTE SET WEIGHTING AND DECOMPOSITION APPROACHES FOR REDUCT COMPUTATION

By

QASEM AHMAD AL-RADAIDEH

July 2005

Chairman : Associate Professor Md. Nasir Sulaiman, PhD

Faculty : Computer Science and Information Technology

This research is mainly in the Rough Set theory based knowledge reduction for data classification within the data mining framework. To facilitate the Rough Set based classification, two main knowledge reduction models are proposed. The first model is an approximate approach for object reducts computation used particularly for the data classification purposes. This approach emphasizes on assigning weights for each attribute in the attributes set. The weights give indication for the importance of an attribute to be considered in the reduct. This proposed approach is named Object Reduct by Attribute Weighting (ORAW). A variation of this approach is proposed to compute full reduct and named Full Reduct by Attribute Weighting (FRAW).

The second proposed approach deals with large datasets particularly with large number of attributes. This approach utilizes the principle of incremental attribute set decomposition to generate an approximate reduct to represent the entire dataset. This proposed approach is termed for Reduct by Attribute Set Decomposition (RASD).

The proposed reduct computation approaches are extensively experimented and evaluated. The evaluation is mainly in two folds: first is to evaluate the proposed approaches as Rough Set based methods where the classification accuracy is used as an evaluation measure. The well known *10-fold* cross validation method is used to estimate the classification accuracy. The second fold is to evaluate the approaches as knowledge reduction methods where the size of the reduct is used as a reduction measure.

The approaches are compared to other reduct computation methods and to other none Rough Set based classification methods. The proposed approaches are applied to various standard domains datasets from the UCI repository. The results of the experiments showed a very good performance for the proposed approaches as classification methods and as knowledge reduction methods. The accuracy of the ORAW approach outperformed the Johnson approach over all the datasets. It also produces better accuracy over the Exhaustive and the Standard Integer Programming (SIP) approaches for the majority of the datasets used in the experiments. For the RASD approach, it is compared to other classification methods and it shows very competitive results in term of classification accuracy and reducts size.

As a conclusion, the proposed approaches have shown competitive and even better accuracy in most tested domains. The experiment results indicate that the proposed approaches as Rough classifiers give good performance across different classification problems and they can be promising methods in solving classification problems. Moreover, the experiments proved that the incremental vertical decomposition framework is an appealing method for knowledge reduction over large datasets within the framework of Rough Set based classification.

Abstrak tesis dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

PENDEKATAN SET ATRIBUT BERPEMBERAT DAN PEMECAHAN BAGI PENGIRAAN PENGURANG

Oleh

QASEM AHMAD AL-RADAIDEH

Julai 2005

Pengerusi : **Profesor Madya Md. Nasir Sulaiman, Ph.D.**

Fakulti: **Sains Komputer dan Teknologi Maklumat**

Penyelidikan ini adalah mengenai pengurangan pengetahuan berasaskan Set Kasar dan pengklasifikasian data dalam kerangka kerja perlombongan data. Bagi memudahkan pengklasifikasian berdasarkan Set Kasar, dua model utama bagi pengurang pengetahuan telah dicadang. Model pertama yang dicadangkan adalah pendekatan anggaran dalam pengiraan objek pengurang yang digunakan khusus untuk tujuan pengklasifikasian data. Pendekatan ini menekankan kepada penggunaan pemberat kepada setiap atribut di dalam set atribut. Pemberat-pemberat ini memberi petunjuk kepada kepentingan sesuatu atribut yang bakal dipertimbangkan di dalam pengurang. Pendekatan ini dinamakan POAB iaitu Pengurang Objek dengan Atribut Berpemberat. Satu variasi kepada pendekatan ini turut dicadangkan bagi mengira pengurang penuh. Variasi ini dinamakan sebagai PPAB bermaksud Pengurang Penuh dengan Atribut Berpemberat.

Model kedua yang dicadangkan melibatkan set data yang besar terutamanya dengan kuantiti atribut yang besar. Pendekatan ini menggunakan prinsip pemecahan set atribut secara berperingkat untuk menjana anggaran pengurang yang mewakili keseluruhan set data. Pendekatan yang dicadangkan ini dinamakan PPST bermaksud Pengurang dengan Pemecahan Set Atribut.

Pendekatan-pendekatan pengiraan pengurang yang dicadangkan dieksperimen dan dinilai secara menyeluruh. Proses penilaian adalah dalam dua aras: pertama adalah penilaian ke atas pendekatan yang dicadangkan berdasarkan Set Kasar di mana ketepatan pengklasifikasian digunakan sebagai ukuran penilaian. Kaedah penilaian bersilang 10-aras yang terkenal juga digunakan bagi menganggar ketepatan pengklasifikasian. Aras kedua penilaian digunakan untuk menilai pendekatan yang dikenali sebagai kaedah pengurang pengetahuan di mana saiz pengurang digunakan sebagai ukuran pengurangan.

Pendekatan-pendekatan ini dibandingkan dengan kaedah pengiraan pengurang yang lain dan termasuk lain-lain ka edah yang tidak berasaskan Set Kasar. Di dalam eksperimen, kami menggunakan pendekatan yang dicadangkan ke atas beberapa set data domain piawai daripada simpanan UCI. Keputusan eksperimen menunjukkan pencapaian yang sangat baik oleh pendekatan yang dicadangkan dalam proses pengklasifikasian dan pengurangan pengetahuan. Ketepatan pendekatan POAB melebihi pendekatan *Johnson* dalam kesemua set data. Ia juga menghasilkan ketepatan yang lebih baik jika dibandingkan dengan pendekatan *Exhaustive* dan *SIP* dalam majoriti set data yang digunakan di dalam eksperimen. Bagi pendekatan PPSA, ianya juga telah dibandingkan dengan kaedah pengklasifikasian yang lain dan telah menunjukkan hasil keputusan yang kompetetif dari segi ketepatan pengklasifikasian dan saiz pengurang yang dijana.

Kesimpulannya, pendekatan-pendekatan yang dicadangkan telah menunjukkan ketepatan yang kompetetif, malah lebih baik apabila diuji menggunakan domain-domain ujian yang utama. Keputusan eksperimen menunjukkan pendekatan pengklasifikasi kasar yang dicadangkan berupaya memberi pencapaian yang baik dan menjanjikan hasil ke atas masalah-masalah pengklasifikasian. Tambahan pula, ekperimen telah membuktikan bahawa kerangkan pemecahan menegak secara berperingkat adalah satu pendekatan yang menarik bagi pengurangan pengetahuan sekiranya menggunakan set data yang besar, dan ianya bernilai untuk digunakan di dalam kerangka pengklasifikasian berasaskan Set Kasar.

ACKNOWLEDGEMENTS

My thanks go firstly, as they should always be, to Allah, who blessed me with the ability to undertake and finally complete this work.

This work could not have been carried out without both direct and indirect help and support from my supervisor, Associate Professor Dr. Md Nasir Sulaiman. Thank you, Dr. Nasir, for introducing me to the topic of Rough Set theory, for support, and for opening doors.

Great thanks go to my supervisory committee members, Associate Professor Hj. Mohd Hasan Selamat and Associate Professor Dr. Hamidah Ibrahim for their valuable comments and fruitful discussions.

This research is partially supported by an IRPA fund. Thanks to University Putra Malaysia and the Malaysian government for the support.

My great thanks go to Yarmouk University, Jordan, for approving the leave to pursue my PhD study. Thanks to all colleagues there for their encouragements.

I will not forget to thank Prof. Dr. T.Y. Lin of San Jose State University, California, USA for his moral support and sharing with me some arguments and perspectives related to Data Mining.

My wife and my lovely daughters, *JOUD* and *DENA*, deserve my unending gratitude for their patience and understanding that make the duration of the study easier for me.

My Mother, Brother, and Sisters have always been a continuous source of support for which I am immensely grateful. They have been patient, encouraging and understanding. My thanks go to my brothers in law and uncles for their encouragements.

I also owe a great debt to my friends, mainly, Mohammad Abu Rahma and Saeed AlHazmi, for their support, for sharing me some enjoyable breaks from my work, and for many interesting conversations.

To everybody else, friends, colleagues, and relatives, thanks for their encouragements, cooperation, and moral support.

Qasem Al-Radaideh

July 2005

TABLE OF CONTENTS

LIST OF TABLES

Table		Page
5.9	Reduct Computation using FRAW Approach	130
5.10	Datasets Accuracy Obtained by RSAD	132
5.11	Comparison of the Accuracy of RASD with FSR Approach	134
5.12	Comparison of the RASD and Zhang Approaches	135
5.13	Comparison of Different Reduct Computation Methods for the Zoo Dataset	136
5.14	Comparison of RASD with the Popular Classification Approaches	137
5.15	Experimenting RASD Approach over the DNA dataset	138
5.16	Comparing RASD Approach with Other Classification Methods over the DNA Dataset	139
5.17	Comparison of RASD and DIFN	141
5.18	Summary Comparison of ORAW and other Classification Approaches	143
A.1	Sample of the **Australian** Credit Card Dataset after Preprocessing	159
A.2	Sample of the **Chess** Dataset after Preprocessing	160
A.3	Sample of the **Cleveland** Dataset after Preprocessing	161
A.4	Sample of the **Glass** Dataset after Preprocessing	161
A.5	Sample of the **Heart Disease** Dataset after Preprocessing	162
A.6	Sample of the **Hepatitis Disease** Dataset after Preprocessing	162
A.7	Sample of the **Iris** Dataset after Preprocessing	163
A.8	Sample of the **Letter** Dataset after Preprocessing	163
A.9	Sample of the **Lymphography** Dataset after Preprocessing	164
A.10.a	Sample of the **Lung Cancer** Dataset after Preprocessing	164
A.10.b	(Continue) Sample of the **Lung Cancer** Dataset after Preprocessing	165
A.11.a	Sample of the **Sonar** Dataset after Preprocessing	166
A.11.b	(Continue) Sample of the **Sonar** Dataset after Preprocessing	167

Table		Page

LIST OF FIGURES

LIST OF ABBREVIATIONS AND NOTATIONS

A	Set of conditional attributes
ACC	Classification ACCuracy
AVCV	Attribute Value Cardinality Vector
$\mathbf{c}_{ij}$	Entry of Discernibility Matrix
CV	Cross Validation
d	The decision attribute in a decision system
DL	Discernibility List
DM	Discernibility Matrix
DMM	Discernibility Matrix Modulo
DS	Decision System
FRAW	Full Reduct by Attribute set Weighting
GAWV	Global Attribute Weight Vector
gw	*g*lobal *w*eight
IND	INDiscernibility relation
IS	Information System
KDD	Knowledge Discovery in Database
LAWV	Local Attribute Wight Vector
lw	*l*ocal *w*eight
OneR	One Rules
ORAW	Object Reducts by Attribute set Weighting
RASD	Reduct by Attribute Set Decomposition
RED	REDuct Set
ROSETTA	ROugh SET Toolkit for Analysis of data
RSES	Rough Set Exploration System
RSESlib	Rough Set Exploration System library
U	Universe of objects
v	An attribute value
Va	The value set of an attribute
vcw	*v*alue *c*ardinality *w*eight

CHAPTER 1

INTRODUCTION

1.1 Background

Due to the explosion of data in our modern society, most organizations have large databases that contain a wealth of undiscovered, yet valuable information. To gain benefits from the collected information and to discover the valuable knowledge, it needs to be analyzed. This leads to a need for methods and ways to aid or substitute humans in the process of knowledge discovery from large datasets. Knowledge discovery and data mining methodologies have been introduced as methods for bridging the knowledge gap between information gathered and information analyzed (Han & Kamber, 2001; Cios *et al.*, 1998; Fayyad *et al.*, 1996b, 1996c). Analogous to the *mining* in the real world, data mining is that, with the computer, we can automatically find the "information gold nuggets" or "diamonds" by sifting out enormous quantities of data-debris from our database.

Data mining is a promising and an interdisciplinary research area spanning several disciplines such as database, machine learning, artificial intelligence, intelligent information systems, statistics, data warehousing and knowledge acquisition in expert systems. Data mining has evolved into an important and active area of research because of theoretical challenges and practical application associated with the problem of discovering interested or previously unknown knowledge from very large real-world databases. With data mining we can simply let data "speak for itself".

There are several tasks in data mining and the most common in the literature is classification, which is a form of data analysis that can be used to extract models describing important data classes. The classification task concentrates on predicting the value of the decision class for an object among a predefined set of classes' values given the values of some given attributes for the object.

In the literature many classification approaches have been proposed and implemented by researchers, such as, decision tree based classification, statistical classification, neural network based classification, genetic algorithms classifiers and Rough Set based classification (Cios *et al.*, 1998; Bazan *et al.*, 2000). Classification has a wide range of applications, including scientific experiments, medical diagnosis, credit approval, etc.

Rough Set theory is a mathematical tool developed as a formal method to turn data into knowledge (Pawlak, 1991). The two main applications of the classical Rough Sets theory are in attribute reduction and classification. Rough Set based classification is inspired by the concepts of the Rough Set theory with a primary goal to extract rules from data represented in a decision system. According to Pawlak (1991), the notion of classification is central to the theory.

A very important issue in data mining is the data redundancy where not all knowledge presented to the data mining task in an information system is necessary to describe it (Pawlak, 1991; Kohavi & Frasca, 1994; Zhang & Yao, 2004; Zhang *et al.*, 2003; Lin & Yin, 2004; Hu *et al.*, 2000; Boussouf & Quafafou, 2001). It is often the case where some of attributes or some of attributes values are superfluous. Rough Sets theory provides the *reduct* concept for data reduction as preprocessing step of data analysis. A reduct is defined as the minimal attribute set preserving classification power of the original information system with the full set of attributes.

The reduct concept of the theory is a fundamental concept towards rule extraction. The concept enables us to discard functionally the redundant information and guarantees that the attributes that do not contribute to the classification are removed. The process of finding reduct is a fundamental step in applying the Rough Set theory for the data classification task. Based on the reduct concept, the rules generated by the classifiers are expected to be more concise than if generated over the original dataset (Pawlak, 1998; Komorowski *et al.*, 1999).

1.2 Problem Statement

Data classification problem is a well known problem in the area of knowledge discovery. In applying Rough Set theory as a classification framework, the problem of computing reducts as a knowledge reduction method, is without doubt the most complex and computer-intensive step in Rough Set data analysis (Pawlak, 1998). The problem of computing all reducts is known to belong to a theoretical class of problems that, informally, requires an amount of computation that grows exponentially with the size of the problem. The problem size is dominated by the number of attributes and objects involved.

Several approximation and heuristic methods have been proposed but there are no universal solutions and no accredited best heuristic method (Kuo & Yajima, 2003). According to Lin & Yin (2004), Kuo & Yajima (2003), and Wang & Chen (2004), so far, the problem of reduct computation stills an open research area in Rough Set theory particularly for large datasets with large number of attributes.

Generally, most of the available heuristic approaches use the discernibility matrix concept and a weighting mechanism to evaluate the significance of the attributes to be considered in the reduct Zhang *et al.* (2003). The available weighting mechanisms may lead to consider some attributes with less importance which eventually lead to low classification accuracy. In addition, some of the available approaches have limitations in handling large amount of datasets particularly with large number of attributes (Bakar *et al*, 2002; Zhengren *et al.*, 2004).

In the available approaches, the most used weight for attributes is the number of occurrences in the discernibility matrix and when several attributes have the same weight a random choice is used. This may allow less significant attributes to be a member of the reduct which lead to low classification accuracy.

Johnson reducer (Nguyen & Nguyen, 1996, Ohrn, 1998) uses the attribute frequency in the discernibility matrix to measure the significance of attributes to be considers in the reduct. A random choice is made when several attributes have the same significance. Hu *et al.* (2000) use the attribute frequency and entry length in discernibility matrix as measures for the significance of attributes. The same